**Improving official statistics with credit and debit card data. Some insights on section bias**

María Valdivieso-Anívarro, Applied Economics, University of Oviedo, [ma.valdivieso@outlook.com](mailto:ma.valdivieso@outlook.com)

María Rosalía Vicente, Applied Economics, University of Oviedo, [mrosalia@uniovi.es](mailto:mrosalia@uniovi.es)

**Abstract**

*This paper explores the use of credit and debit card transactions to measure consumption and poverty, the latter understood as the absence of consumption or the restriction of it. These data are available on a timely basis and provide detailed information about spending activity with a finer granularity than that available in traditional household surveys. Accordingly, the exploitation of these data is claimed to provide novel insights about both consumption patterns and poverty. Nonetheless, it is essential not to neglect potential selection biases. Using individual data from credit and debit card payments, we evaluate the potential of these data to estimate consumers’ spending activity. We first investigate selection biases. By looking at credit card users’ self-reported incomes, results show that recorded transactions basically refer to middle/upper incomes. In fact, our data appear to be biased towards the top 50% of the population. We also analyze the spatial distribution of transaction records by city district and compare them with figures derived from the application of small area estimation techniques. Results show that, despite the identified sample selection bias, there are fairly high correlations between self-reported customers’ incomes and monetary poverty rates at district level.*

**Keywords:** Credit/debit card transactions, consumption, poverty, nowcasting

**1. Introduction[[1]](#footnote-2)**

In recent years, a rich body of empirical literature has analyzed the potential of using Big Data to improve official statistics. Several types of data sources have been explored, the most popular being the volume of keywords sought on online search engines (mainly Google), the number of visits to websites (e.g. Wikipedia) and posts on online social networks (Facebook, Twitter, and Instagram, among others). Research using credit card or mobile phone records is comparatively less since these data are essential for banks and telecom companies’ business strategies[[2]](#footnote-3), and consequently, they are little willing to share them (Klein and Verhulst, 2017).

The use of credit and debit card transaction records is claimed to offer some advantages over traditional survey-based statistics for the estimation of private consumption and spending: while the latter statistics depend on respondents’ accurately recalling what they have purchased and how much they have spent; the former data do not suffer from these potential measurement errors (Duarte et al., 2017). In addition, these data are available on a timely basis and provide very detailed information about spending activity (when, where, what) with much lower costs compared to survey data collection.

Some previous research has shown that credit and debit card transaction records can help to improve the estimation of some macroeconomic variables such as Gross Domestic Product, retail sales and private consumption in high income countries (Buono et al., 2017and the references cited therein; Galbraith and Tkacz, 2007, 2015; Duarte et al., 2017; Gil et al., 2017). There are also a couple of recent applications for middle-low income countries: Di Clemente et al. (2018) and Vaitla et al. (2017) combine data on credit card transactions with mobile phone records to study habits of consumption, spending priorities and shopping patterns in one the largest cities in Latin America and derive some proxies for lifestyles.

While the granularity of these data is especially attractive to get detailed insights about individual economic behaviour (and potentially to improve the nowcasting of socioeconomic indicators), it is essential not to neglect the disadvantages associated with this kind of data. Importantly, credit and debit card transactions-related data suffer from (self-)selection bias: neither do they represent the full population (i.e., only individuals who have a credit/debit card) or characterize all private consumption (i.e., only consumption paid with credit or debit cards). Such selection biases might be particularly worrying in the case of middle-low income countries, where the credit card usage tends to be low. Di Clemente et al. (2018) and Vaitla et al. (2017) report that less than a quarter of the population use credit card in the areas of analysis. In particular, Vaitla et al. (2017, p.18) explain that “(within neighborhoods) the sample is biased towards wealthier individuals—although, evaluating the sample as a whole, users from all income levels are well-represented.”

This paper attempts to provide further insights on the potential usefulness of credit and debit card transaction records to improve the nowcasting of official statistics and, in particular, private consumption indicators. To achieve this, we analyze individual data from customers of one of the largest banks in Latin America, paying special attention to selection biases.

**2. Data**

Data include anonymized individual credit and debit card records from customers of one of the largest banks in Latin America. The dataset contains transaction records for a full year (from June 2016 through May 2017) from bank’s customer base in one of the biggest countries where it operates[[3]](#footnote-4).

For each transaction, we have detailed information about the date when it took place, the district of the shop where the purchase was done, the total amount spent as well as broken down by expenditure group, considering 35 categories (e.g., Grocery Stores, Supermarkets; Eating Places, Restaurants; Service Stations; Drug Stores and Pharmacies; Tax Payments - Government Agencies; and Book Stores, among others). Each transaction is also associated with customer’s id, along with some socioeconomic features, i.e., age, income, gender, and residence (district and province). Table 1 shows an example of how the database looks like.

**Table 1. Sample credit/debit card detail records**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Client** | **Date** | **District** | **Province** | **Age** | **Income** | **Sex** | **Amount** | **Shop\_District** | **Exp\_520** | **Exp\_521** |
| Aaaa | 24/02/2017 | district 1 | province 1 | 28 | 1691 | M | 20 | shop\_d1 | 0 | 20 |
| Bbbb | 21/10/2016 | district 2 | province 2 | 50 | 3341 | F | 11 | shop\_d2 | 11 | 0 |
| Cccc | 14/04/2017 | district 3 | province 3 | 61 | 2250 | F | 89 | shop\_d3 | 89 | 0 |
| Dddd | 20/06/2016 | district 4 | province 4 | 36 | 5258 | M | 148 | shop\_d4 | 0 | 148 |

Source:Credit/debit card database.

While data are available for the full country, our analysis has focused on customers living in the capital city where the usage of credit and debit cards is the highest. In this way, we have tried to limit as much as possible potential biases derived from low credit card penetration rates in rural areas.

**3. Results**

In order to identify selection biases in our data, we first analyze the incomes of credit card users and compare them with official figures. Table 2 shows the deciles of self-reported monthly individual incomes in our database. The first decile is 853, that is, the 10 percent of credit card users with the lowest reported incomes have less than 853 monetary units per month. According to official statistics (national household survey)[[4]](#footnote-5), the corresponding figure (first monthly income decile) is around 320monetary units per capita in the area of analysis. Moreover, official figures report that half of the population in the capital city gets less than 850 per month. Consequently, our data appear to be biased towards the upper 50 percent. In fact, the richest decile in our database gets over 7,700 monetary units per month, which almost doubles the figure reported by official statistics for the top income decile.

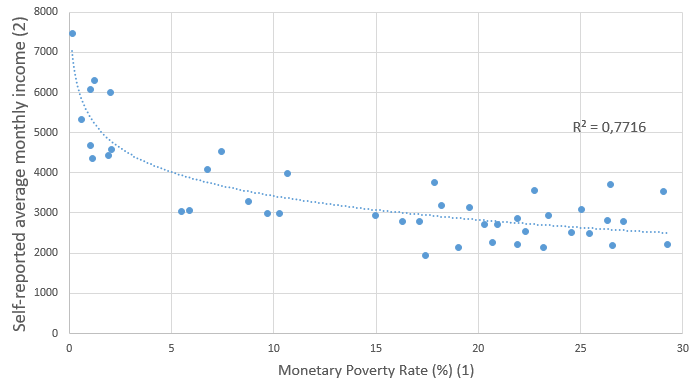
**Table 2. Self-reported average monthly incomes in capital city in 2016**

|  |  |
| --- | --- |
| **Decile** | **Self-reported monthly individual income** |
| 1st | 853 |
| 2nd | 1200 |
| 3rd | 1525.63 |
| 4th | 1880.5 |
| 5th | 2314.22 |
| 6th | 2912.56 |
| 7th | 3811.33 |
| 8th | 5185.67 |
| 9th | 7753 |

Since our data are detailed at city district, we have tried to get further insights into the identified selection bias. In particular, we have calculated credit card users’ self-reported average (and median) monthly incomes by capital city districts. The problem here is that income-related official statistics are not disaggregated at this geographical level. The only available statistics refer to the estimates of monetary poverty rates which are derived from the application of small area estimation techniques. The monetary poverty rate is defined as the percentage of people whose expenditure is below the basic consumption basket (including food and non-food items). Intuitively, we can expect that (i) most of our data observations come from districts with low poverty rates and (ii) poverty rates will be negatively correlated with self-reported average (median) monthly incomes by city districts. Results show that the number of records available at district level is negatively correlated with poverty rates. Specifically, we have calculated the linear correlation coefficients between both the number of credit card users and the number of transaction records by district and poverty rates, getting values in all cases around -0.4.

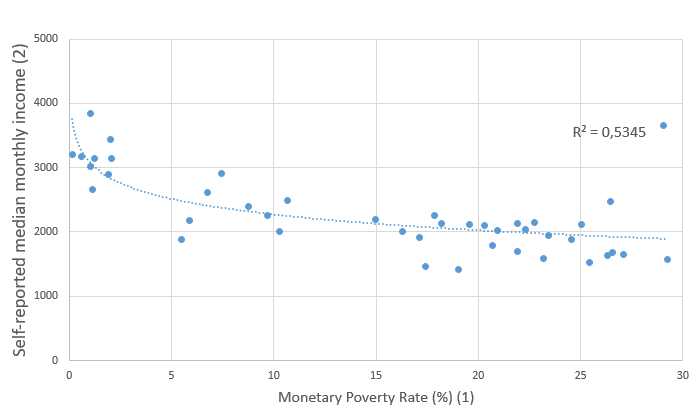
Figures 1 and 2 show the scatter plots for the monetary poverty rates and the self-reported average and median monthly incomes of credit cards users at district level. Both plots show a negative relationship between poverty rates and self-reported incomes, with linear correlation coefficients between -0.6 and -0.75. This relationship can be fitted through the estimation of exponential models in which district-related poverty rates are explained by self-reported average and median incomes with R2 of 0.77 and 0.53, respectively.

**Figure 1. Monetary poverty rates and self-reported average monthly incomes (2016) by capital city districts**

****

Notes: (1) Monetary poverty rates are defined as % of people whose expenditure is below the basic consumption basket. (2) Credit/debit card records database

**Figure 2. Monetary poverty rates and self-reported median monthly incomes (2016) by capital city districts**



Notes: See notes under Figure 1.

**4. Concluding remarks**

Recent research on the potential use of big data to improve the nowcasting of socioeconomic indicators has usually emphasized the advantages of these data sources in terms of the volume, granularity and timely availability of data. However, the use of these data is not without problems, especially in what refers to selection biases. While applications of big data sources generally recognize the possible existence of this kind of problem, there is a lack of in-depth analysis of the extent of these potential selection biases. As Hartford (2014) indicates: “big data’ has arrived, but big insights have not. The challenge now is to solve new problems and gain new answers - without making the same old statistical mistakes on a grander scale than ever.”

This paper has tried to show some evidence on this issue by analyzing individual data from credit and debit card payments and comparing it with survey-based statistics. Results show that our database (as a whole) is biased towards the top 50% of the population. When data are disaggregated by city district, negative correlations are found between poverty rates and both the number of records and self-reported customer incomes. Despite the identified selection biases, the fairly high correlations observed at district level suggest that the combination of these data with other big data sources along with other socioeconomic statistics at district level might help to improve indicators derived from small area estimation techniques as shown by Marchetti et al. (2015, 2017). Future research of this ongoing project will explore such issues.

**5. References**

Buono, D., Mazzi, G.L, Kapetanios, G., et al. (2017), Big Data Types for Macroeconomic Nowcasting (pp.93-145). In Eurostat (Ed.), Eurostat Review on National and Macroeconomic Indicators (1/2017), Eurostat, Brussels.

Di Clemente, R., Luengo-Oroz, M., Travizano, M. et al. (2018), Sequence of Purchases in Credit Card Data Reveal Life Styles in Urban Population. Available at: <http://www.mit.edu/~rdicle/spending.html> (Accessed: 30 April 2018).

Duarte, C., Rodrigues, P.M.M., and Rua, A. (2017),A Mixed Frequency Approach to the Forecasting of Private Consumption with ATM/POS Data, International Journal of Forecasting, 33(1), pp. 61-75.

Galbraith, J.W and Tkacz, G.(2007), Analyzing Economic Effects of Extreme Events using Debit and Payments System Data, CIRANO Scientific Series, Working Paper 2011s-70.

Galbraith, J.W. and Tkacz, G. (2015), Nowcasting GDP with Electronic Payments Data, European Central Bank, Working Paper No 10 / August 2015

Gil, M., Pérez, J.J, and Urtasun, A. (2017), Nowcasting Private Consumption: Traditional Indicators, Uncertainty Measures, and the Role of Internet Search Query Data. Available at: <https://ecomod.net/system/files/Urtasun_Consumption_and_big_data.pdf>(Accessed: 26 April 2018).

Hartford, T. (2014), Big Data: Are We Making a Big Mistake?, Financial Times, 28.03.2014. Available at: <https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0> (Accessed: 26 April 2018).

Hernández, M., Hong, L., Frias-Martínez, V. et al. (2017), Estimating Poverty Using Cell Phone Data. Evidence from Guatemala, World Bank Working Paper, 7969.

Klein, T., andVerhulst, S. (2017), Access to New Data Sources for Statistics: Business Models and Incentives for the Corporate Sector, OECD Statistics Working Papers (2017/06), OECD, Paris.

Marchetti, S., Giusti, C., and Pratesi, M. (2017). The Use of Big Data from Twitter for Small Area Estimation of Households' Share of Food Consumption Expenditure in Italy (pp.44-58). In C. Articus et al. (Eds.), Future Needs in Statistics, InGRID Deliverable 23.2, InGRID project, Trier.

Marchetti, S., Giusti, C., Pratesi, M., et al. (2015), Small Area Model-Based Estimators Using Big Data Sources, Journal of Official Statistics, 31(2), pp. 263-281.

Vaitla, B., Bosco, C., Alegana, V. et al. (2017), Big Data and The Well-Being of Women and Girls, Report of Data2X with the UN Foundation. Available at: http://humnetlab.mit.edu/wordpress/publications/ (Accessed: 26 February 2018).

1. This paper is part on an ongoing project about the use of big data sources to improve official statistics in middle/low income countries. [↑](#footnote-ref-2)
2. See Hernández et al. (2017) for a recent application of cell phone data to estimate poverty rates in Guatemala. [↑](#footnote-ref-3)
3. Names are omitted for confidentiality reasons. [↑](#footnote-ref-4)
4. As stated in the previous footnote, specific names are omitted for confidentiality reasons. [↑](#footnote-ref-5)